

Significance Analysis of High-Dimensional, Low-Sample Size Partially Labeled Data

Qiyi Lu[†] and Xingye Qiao

Department of Mathematical Sciences
Binghamton University, State University of New York
Binghamton, New York, 13902-6000

E-mails: {qlu,qiao}@math.binghamton.edu

Phone: (607) 777-2147

Fax: (607) 777-2450

September 22, 2015

[†]Correspondence to: Qiyi Lu (e-mail: qlu@math.binghamton.edu). Qiyi Lu is a doctoral candidate and Xingye Qiao (e-mail: qiao@math.binghamton.edu) is an Assistant Professor at Department of Mathematical Sciences at Binghamton University, State University of New York, Binghamton, New York, 13902-6000. Qiao's research is partially supported by a collaboration grant from *Simons Foundation* (award number 246649).

Abstract

Classification and clustering are both important topics in statistical learning. A natural question herein is whether predefined classes are really different from one another, or whether clusters are really there. Specifically, we may be interested in knowing whether the two classes defined by some class labels (when they are provided), or the two clusters tagged by a clustering algorithm (where class labels are not provided), are from the same underlying distribution. Although both are challenging questions for the high-dimensional, low-sample size data, there has been some recent development for both. However, when it is costly to manually place labels on observations, it is often that only a small portion of the class labels is available. In this article, we propose a significance analysis approach for such type of data, namely partially labeled data. Our method makes use of the whole data and tries to test the class difference as if all the labels were observed. Compared to a testing method that ignores the label information, our method provides a greater power, meanwhile, maintaining the size, illustrated by a comprehensive simulation study. Theoretical properties of the proposed method are studied with emphasis on the high-dimensional, low-sample size setting. Our simulated examples help to understand when and how the information extracted from the labeled data can be effective. A real data example further illustrates the usefulness of the proposed method.

KEY WORDS: Classification; Clustering; High-dimensional, low-sample size data; Hypothesis test; Semi-supervised learning.

1 Introduction

Classification and clustering are both important tools in statistical learning. The availability of the class labels distinguish these two main domains. In classification, class labels are provided prior to the analysis, while they are unavailable in the clustering analysis. A natural statistical question regarding their use is whether classes/clusters are really there. In a setting where binary class labels are observed, we may be interested in testing whether the two classes are from the same distribution. Though often neglected, this is an important step before applying a classification algorithm. In standard statistical textbooks, there are many significance tests, such as two-sample t -test, one-way ANOVA, Hotelling's T^2 test, and MANOVA. Among these, the two-sample t -test and ANOVA are univariate tests. The Hotelling's T^2 test and MANOVA are multivariate tests, though both can fail when the dimension d is much greater than the sample size n .

This problem of testing the difference between two classes becomes even more challenging for the high-dimensional, low-sample size (HDLSS) data. The Hotelling's T^2 test is very powerful when the dimension is smaller than the sample size. It is invariant under linear transformation. In addition, under the null hypothesis, the distribution of the statistic is known. However, the Hotelling's T^2 statistic cannot be computed in the HDLSS setting because the sample covariance matrix is not invertible. There are efforts attempting to overcome this issue, including [Dempster \(1960\)](#), [Bai and Saranadasa \(1996\)](#), [Srivastava and Du \(2008\)](#) and [Chen and Qin \(2010\)](#). These methods use diagonalized versions of the covariance or inverse covariance matrices in the Hotelling's T^2 statistic. There are many other treatments, such as [Srivastava and Fujikoshi \(2006\)](#), [Schott \(2007\)](#) and [Srivastava \(2007\)](#), which calibrate the distribution of some proposed statistic. In addition, the Direction-Projection-Permutation (DiProPerm) test ([Wichers et al., 2007](#), [Wei et al., 2015](#)) has been proved to be very effective for testing the class difference of the HDLSS data.

Besides the difficulty brought from the high dimensionality, in many real problems, it is often the case that there are many observations that are left without class labels (the unlabeled data portion) in a data set. One reason is that it is often difficult or expensive to

obtain the class label information, while it may be relatively cheap to obtain the covariate information even for many observations. In such a situation, those aforementioned testing methods which require label information cannot be applied to the whole data set to test the class difference. As a consequence, one may have to forfeit the potentially useful information that resides in the unlabeled data. For instance, many cancer patients are categorized to certain cancer subtypes by radiologists through an inspection of the medical images. However, because of the high health care cost, medical images are easier to obtain than the actual diagnostic. Before a classification algorithm is used to design a data-mining-based early-screening machine (see, for example, [Land et al., 2012](#), [Schaffer et al., 2012](#)), a valuable question to ask is whether the so-called subtypes, many of which may be ad hoc or based on experience, are really there.

One possible, but clearly flawed, solution to this problem is to treat all the data as unlabeled. In the unsupervised context, in the sense that there are no class labels provided for the analysis, clustering algorithms have been broadly applied in many fields. As to determine whether clusters are really there, several methods have been developed to assess the significance of clusters, including [McShane et al. \(2002\)](#), [Tibshirani and Walther \(2005\)](#), [Suzuki and Shimodaira \(2006\)](#) and [Liu et al. \(2008\)](#). However, these methods are not directly applicable for partially labeled data, unless one forfeits the potentially useful information that resides in the class labels which are available in the labeled data portion of the full data set.

Hence, there seems to be a dilemma in testing partially labeled data: to ignore the unlabeled data completely (and apply a significance test for the labeled data only), versus, to ignore the class labels in the labeled data portion (and use a significance test for clustering). Although each has its own applicability domain, neither looks promising for us. This motivates us to devise a significance testing method for the HDLSS partially labeled data. When class labels are partially provided, the unlabeled data are used to better estimate the sampling distribution. In the meantime, the class labels help to effectively distinguish the two classes even if their difference is small. Our proposed method is named Significance

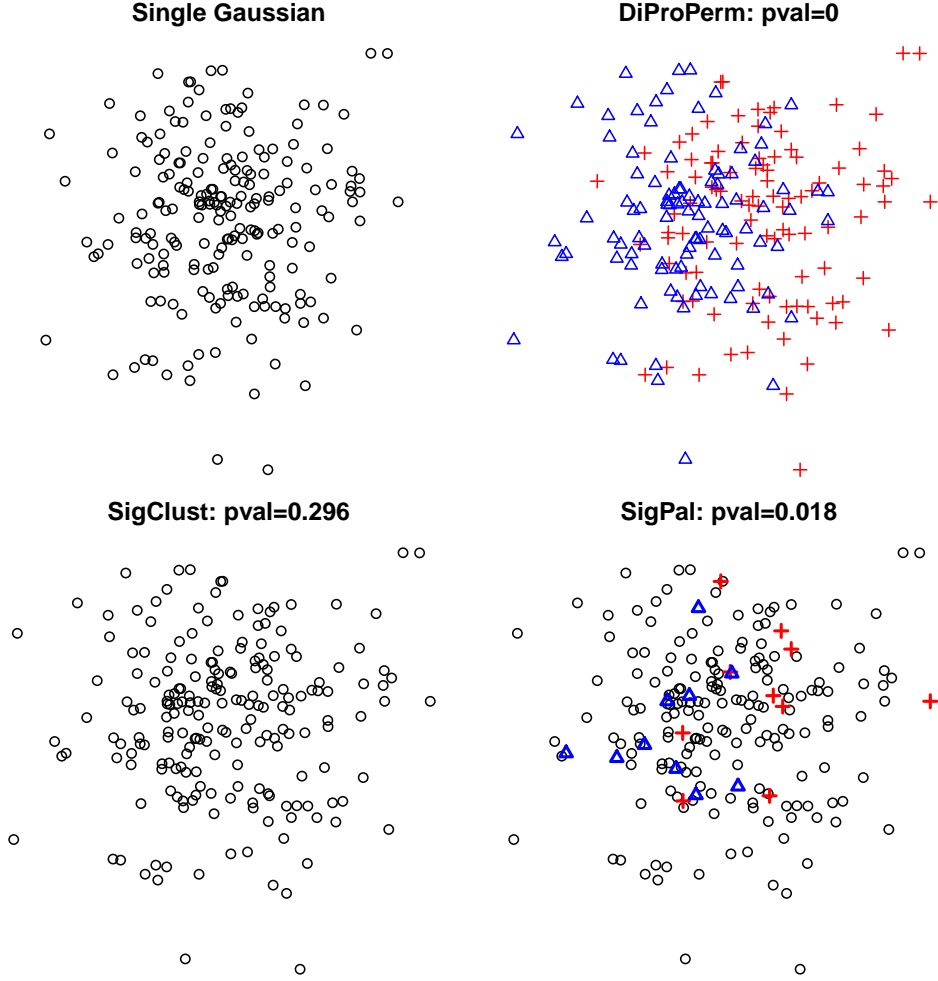


Figure 1: The DiProPerm test is applicable when all class labels are known (top-right panel) while SigClust does not require any label information (bottom-left panel). The DiProPerm test correctly concludes that the two classes are indeed from two distributions (with p -value= 0), whereas the SigClust method fails to find this important difference (p -value= 0.296). When the majority of the data are unlabeled with a small portion of labeled data, our proposed SigPal approach can give a significant conclusion with p -value= 0.018, which is close to the DiProPerm result.

To illustrate our main idea, we show a toy example under different settings in Figure 1. The data in the top-left panel are generated from a Gaussian distribution and the data in the rest three panels come from a mixture of two Gaussian distributions with a small difference in the mean, $0.5N(-\boldsymbol{\mu}, \mathbf{I}_2) + 0.5N(\boldsymbol{\mu}, \mathbf{I}_2)$, where $\boldsymbol{\mu} = (0.5, 0)'$. To ease the presentation,

the toy example is of two-dimensional, though the message applies to the HDLSS data. We show two significance analysis methods for the HDLSS data that inspire our approach, the DiProPerm test of [Wichers et al. \(2007\)](#) and [Wei et al. \(2015\)](#), and the Statistical Significance of Clustering method (SigClust) of [Liu et al. \(2008\)](#) and [Huang et al. \(2014\)](#). The DiProPerm test is applicable when all the class labels are known (see the different colors/marker-types in the top-right panel, where each component of the mixture distribution corresponds to one class.) In contrast, SigClust does not require any label information (bottom-left panel). The (empirical) p -value of the DiProPerm test turns out to be 0, which leads to a correct conclusion that the two classes are indeed from two distributions, whereas the SigClust method fails to find this important difference (p -value= 0.296). Our proposed SigPal approach is designed for the case shown in the bottom-right panel. Given some labeled data, SigPal can give a significant conclusion with p -value= 0.018, which is close to the DiProPerm result. All these three methods will be introduced or revisited in the next two sections.

The rest of the article is organized as follows. In Section 2, we review the DiProPerm test and the SigClust method. Section 3 presents our proposed SigPal method. Some theoretical results are studied in Section 4 which emphasize the HDLSS setting. A comprehensive simulation study and real data case study are provided in Section 5. Section 6 gives some concluding remarks. The appendix is devoted to technical proofs.

2 DiProPerm Test and SigClust Test

In this section, we review two significance analysis methods, DiProPerm and SigClust. Both methods are specifically designed for testing HDLSS data, although they may be applied to low-dimensional data as well. DiProPerm is used when all the class labels are fully observed while SigClust is applicable when the data set has no class label information. The hypotheses of both methods are slightly different.

2.1 DiProPerm Test

In practice, permutation tests are often used for the purpose of testing the class difference, where the null distribution is mimicked by the empirical distribution of the statistic calculated from many randomly permuted data sets. However, for high-dimensional data, some distance measure with direct permutation may not work. This is because when $d \gg n$, such distance measure will be mainly driven by the error aggregated over dimensions, rather than the true mean difference between classes. To address this issue, a three-step procedure called Direction-PROjection-PERmutation test (DiProPerm) was proposed in [Wichers et al. \(2007\)](#) and further studied in [Wei et al. \(2015\)](#) for the two-class setting. DiProPerm was designed for data with fully observed labels. It tests the null hypothesis of equality of distributions:

H_0 : the distributions of the two classes are the same, and

H_1 : the distributions of the two classes are not the same.

Another item of interest is to test the weaker null hypothesis of equality of means:

H_0 : the distributions of the two classes have equal means, and

H_0 : the distributions of the two classes have different means.

ALGORITHM 1. The DiProPerm test comprises three steps.

1. **Direction:** a direction which is capable of separating the two classes is found, such as the classification direction from Support Vector Machine (SVM; [Vapnik, 1995](#), [Cortes and Vapnik, 1995](#)), Distance Weighted Discrimination (DWD; [Marron et al., 2007](#)), or their hybrids ([Qiao and Zhang, 2015b,a](#)).
2. **Projection:** all the data vectors are projected to this direction so that a univariate statistic (such as the two-sample t -statistic or the mean difference) can be calculated.
3. **Permutation:** all the data are randomly relabeled and the first two steps are repeated for N_{Perm} times (N_{Perm} may be 1000.) An empirical p -value is calculated to assess the statistical significance (the proportion of the statistics from the permutation set that are greater than that from the data).

In Figure 2, we illustrate how the p -value in DiProPerm is calculated, using the same data as shown in Figure 1. In the left plot, we perform a DiProPerm test with 1000 permutations. The test statistics calculated for the permutations are shown as the blue jitter points, while that for the original data is the green vertical line. Here the mean difference is chosen as the statistic. The greater the statistic is, the more significantly different the two classes are. Hence, the empirical p -value is calculated as the proportion of the statistics from the permutation set that are greater than that from the data, which is 0 in this case.

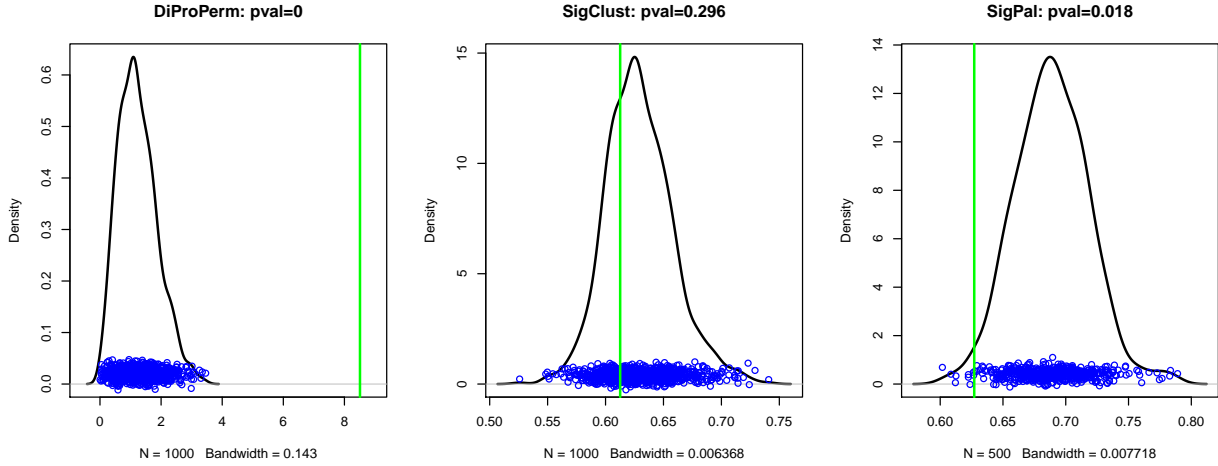


Figure 2: Illustration of the calculation of the p -values for DiProPerm (left), SigClust (middle) and SigPal (right). The test statistics for the permutation/simulation set are shown as blue jitter points, while those for the original data are shown as the green vertical lines. The empirical p -value for DiProPerm is the proportion of the statistics from the permutation set that are greater than that from the data, which is 0 in this case. The empirical p -values for both SigClust and SigPal are the proportions of the statistics from the simulation set that are less than that from the data, which are 0.296 and 0.018 respectively.

The main idea of DiProPerm is to measure the difference between two high-dimensional data subsets by the difference between their 1-dimensional projections onto some appropriate direction. DiProPerm is a powerful test in many settings and it is a nonparametric procedure that does not have many assumptions. Note that here class labels are required to find the projection direction (by a classification method such as SVM or DWD) and to calculate the test statistic (t -statistic or mean difference).

2.2 SigClust

SigClust, proposed by [Liu et al. \(2008\)](#) and improved by [Huang et al. \(2014\)](#), is a clustering evaluation tool for the HDLSS data which aims to answer the question whether clusters are really there. That is, it has the following hypotheses:

H_0 : the data are from a single Gaussian distribution, and

H_1 : the data are not from a single Gaussian distribution.

There is no reference to the notion of class in the hypotheses above. SigClust is based on the vision of cluster as a subset of the data that can be reasonably modeled as coming from a single Gaussian distribution (with some covariance matrix). The Gaussian assumption has been previously used by [Sarle and Kuo \(1993\)](#) and [McLachlan and Peel \(2004\)](#). [Huang et al. \(2014\)](#) mentioned that this assumption may lead to some important consequences. For example, it is possible that none of Cauchy, Uniform, or even t distributed data may be viewed as a single cluster in this sense. While it may seem to be a strong assumption, it is reasonable in the challenging HDLSS situation because it allows real HDLSS data analysis with wide use in bioinformatics applications ([Chandriani et al., 2009](#), [Verhaak et al., 2010](#)).

Assume that a data set $\{\mathbf{x}_i, i = 1, \dots, n\}$ is obtained from an unknown Gaussian distribution with covariance Σ , where $\mathbf{x}_i \in \mathbb{R}^d$ is the observed covariates. The idea of SigClust is to approximate the null distribution of a test statistic by simulating from a single Gaussian distribution that fits to the data. The p -value in SigClust is taken to be the lower quantile of the null distribution, defined by the test statistic from the original data. It is similar to the DiProPerm test except that it performs simulation instead of permutation and it relies on a multivariate statistic instead of a univariate statistic after projection.

Specifically, the 2-means Cluster Index (CI) is used as the test statistic. It is defined as the sum of the within-cluster sums of squares about the cluster means, divided by the sum of squares about the overall mean,

$$CI = \frac{\sum_{k=1}^2 \sum_{j \in C_k} \|\mathbf{x}_j - \bar{\mathbf{x}}^{(k)}\|^2}{\sum_{j=1}^n \|\mathbf{x}_j - \bar{\mathbf{x}}\|^2},$$

where C_k denotes the sample index set of the k th cluster and $\bar{\mathbf{x}}^{(k)}$ represents the mean of

the k th cluster, for $k = 1, 2$. The smaller the CI, the larger the proportion of the overall variation that is explained by the clusters. Note that no predefined class labels are needed when computing the CI, as the cluster assignment C_k is obtained concurrently by a clustering algorithm.

Here the simulation from the null distribution is a critical part. As CI is location and rotation invariant, it is enough to work only with a Gaussian null distribution with a mean at the origin and a diagonal covariance matrix Λ . Hence, an essential part of the SigClust test is the estimation of the eigenvalues of the covariance matrix Σ .

ALGORITHM 2. The SigClust procedure is summarized as follows.

1. **Initialization:** obtain a two-cluster assignment ($k = 2$) from an application of a clustering algorithm, such as k -means. The CI is then calculated for the original data set based on the cluster assignment.
2. **Simulation:** simulate data from the null distribution: (X_1, \dots, X_d) are independent with $X_j \sim N(0, \hat{\lambda}_j)$, where $(\hat{\lambda}_1, \dots, \hat{\lambda}_d)$ is an estimate of the eigenvalues $(\lambda_1, \dots, \lambda_d)$ of the covariance matrix Σ . Then calculate the corresponding CI on each simulated data after performing clustering in the same manner as in the **Initialization** step.
3. **Testing:** repeat the **Simulation** step for N_{Sim} times to obtain an empirical distribution of CI based on the null hypothesis (N_{Sim} may be 1000). Then calculate the empirical p -value to assess the statistical significance (the proportion of the CI from the simulation set that are less than the CI from the original data.)

The middle plot in Figure 2 illustrates how the p -value in SigClust is calculated. Similar to the DiProPerm case (left plot), the blue jitter plots are the statistics from the simulations, and the green vertical line is that for the original data. Recall that the smaller the statistic (chosen as the CI) is, the more significantly different the two clusters are. Hence the empirical p -value is the proportion of the CI's from the simulation set that are less than that from the original data, which is 0.296 in this case.

The covariance estimation in the **Simulation** step can be challenging, especially when the data are HDLSS. Although we only need to estimate the eigenvalues of the covariance matrix, which greatly reduces the number of parameters to be estimated, this problem is still not trivial in the HDLSS setting. [Liu et al. \(2008\)](#) used a hard-thresholding approach for eigenvalue estimation. In particular, they first estimate the background noise level using a robust variance estimate. Then those estimated eigenvalues smaller than the background noise level are replaced with the noise level, that is,

$$\hat{\lambda}_j = \begin{cases} \tilde{\lambda}_j & \text{if } \tilde{\lambda}_j \geq \hat{\sigma}_N^2 \\ \hat{\sigma}_N^2 & \text{if } \tilde{\lambda}_j < \hat{\sigma}_N^2 \end{cases},$$

where $(\tilde{\lambda}_1, \dots, \tilde{\lambda}_d)$ are the eigenvalues of the sample covariance matrix and $\hat{\sigma}_N^2$ is the estimated background noise level.

[Huang et al. \(2014\)](#) later showed that with eigenvalue estimation using hard-thresholding, SigClust can be seriously anti-conservative if the first eigenvalue is relatively large. They proposed a less-aggressive soft-thresholding variant which greatly improved the performance of SigClust. Specifically, they use

$$\hat{\lambda}_j = \begin{cases} \tilde{\lambda}_j - \tau & \text{if } \tilde{\lambda}_j \geq \tau + \hat{\sigma}_N^2 \\ \hat{\sigma}_N^2 & \text{if } \tilde{\lambda}_j < \tau + \hat{\sigma}_N^2 \end{cases}.$$

A detailed definition of τ can be found in [Huang et al. \(2014\)](#).

3 Significance Analysis for Partially Labeled Data

In this section, we first state the background and hypotheses of our problem, followed by a presentation of our proposed method.

3.1 Background and Hypotheses

Consider a binary testing problem for a data set with the labeled data portion $\{(\mathbf{x}_i, y_i), i = 1, \dots, n_l\}$, and the unlabeled data portion $\{\mathbf{x}_{n_l+j}, j = 1, \dots, n_u\}$. All the \mathbf{x}_i 's and \mathbf{x}_{n_l+j} 's

are d -dimensional covariates and the class label $y_i \in \{-1, 1\}$. The total sample size is $n = n_l + n_u$. Let $\theta = n_l/n$ be the proportion of the labeled data, and $1 - \theta$ is the proportion of the unlabeled data. We formulate our proposed SigPal procedure as a hypothesis testing problem with the following hypotheses:

H_0 : the data come from a single Gaussian distribution, and

H_1 : the conditional distributions of the two classes are different and hence the data are not from a single Gaussian distribution.

It is worth comparing our alternative hypothesis with those of the DiProPerm and SigClust tests. Since not every class label is observed, the notion of the class, as in the alternative hypothesis of DiProPerm, is moot or murky. Technically, our alternative hypothesis is neither an intersection nor a union of the previous alternative hypotheses. In the framework of SigPal, there exists an underlying class label for each observation. We are interested in the difference in the conditional distributions of the data with respect to this underlying label. Our goal is to infer the significance of the otherwise fully observed data based on the partially labeled data with the help of the covariate information of the whole data. Lastly, the fact that the conditional distributions are different implies that the data are not from a single Gaussian distribution (but the converse is not true.)

3.2 Proposed Method

With different values of the proportion of the labeled data, θ , we may consider different ways to address the problem. When θ is close to 1, which means the majority of the data have label information available, then one may just ignore the small amount of unlabeled data and perform a DiProPerm test on the labeled data portion only. When θ is very close to 0, which means the majority of the data are unlabeled, then one can simply apply SigClust regardless of the few class labels. While we may lose some useful information from the data, such simplifications effectively reduce the complexity of the problem. In this article, we are more interested in the case when θ is not close to 0 or 1. We propose a Significance Analysis for Partially Labeled Data (SigPal) which makes use of the extra label information,

compared to the SigClust procedure.

ALGORITHM 3. The SigPal procedure consists of the following three steps.

1. **Initialization:** obtain the predicted class assignments for the unlabeled data by applying a semi-supervised classification/clustering method to the full data set. A test statistic is then calculated for the original data based on both observed and predicted class labels (our choice is CI in this article).
2. **Simulation:** simulate data from the null distribution: (X_1, \dots, X_d) are independent with $X_j \sim N(0, \hat{\lambda}_j)$, where $(\hat{\lambda}_1, \dots, \hat{\lambda}_d)$ is an estimate of the eigenvalues $(\lambda_1, \dots, \lambda_d)$ of the covariance matrix Σ . Randomly place class labels to n_l observations in the simulated set and then calculate the corresponding test statistic after performing semi-supervised classification/clustering in the same manner as in the **Initialization** step.
3. **Testing:** repeat the **Simulation** step for N_{Sim} times to obtain an empirical distribution of the test statistic based on the null hypothesis. Calculate the empirical p -value (the proportion of the CI's from the simulated data that are less than that from the original data) to assess the statistical significance.

The right plot in Figure 2 shows the p -value calculation for SigPal. Similar to the SigClust, the empirical p -value is the proportion of the CI's from the simulation set that are less than that from the original data, which is 0.018 in this case.

To calculate the statistic in the **Initialization** step, we need to assign labels for the unlabeled portion. This is similar to the application of a clustering algorithm in SigClust. Such label assignment can be done either by modifying a classification method or by modifying a clustering algorithm.

- While we could simply use a classifier trained from the labeled portion to predict the class label for the unlabeled portion, a semi-supervised classification method is more reasonable here since it takes the covariate information in the large number of unlabeled observations into account. Possible choices of the semi-supervised classification method

include Semi-Supervised Sparse Linear Discriminant Analysis (S^3 LDA; [Lu and Qiao, 2015](#)), transductive SVM (TSVM; [Vapnik, 1998](#), [Chapelle et al., 2006](#), [Wang et al., 2007](#)), the large-margin based methods ([Wang and Shen, 2007](#), [Wang et al., 2009](#)) and the bootstrap method ([Collins and Singer, 1999](#)). S^3 LDA combines the classical linear discriminant analysis and a machine learning oriented technique, and takes advantage of the unlabeled data to boost the classification performance.

- Similarly, though we could just run a clustering algorithm for the whole data set to assign labels, we would like to borrow the strength in the labeled data portion. A semi-supervised clustering algorithm, which identifies clusters with constraints imposed by known labels, would be more appropriate in this case. Possible semi-supervised clustering algorithms include constrained k -means (COP-KMEANS; [Wagstaff et al., 2001](#)), and others. COP-KMEANS allows a must-link constraint which specifies that certain instances have to be placed in the same cluster.

Once the class/cluster labels are assigned, a test statistic can be calculated. Options include the Hotelling’s T^2 statistic, the CI, and some one-dimensional statistics (such as two-sample t -statistic or mean difference) after projections as in DiProPerm. CI is more favorable here since it is location and rotation invariant and can be efficiently computed. It also facilitates the comparison between SigPal and SigClust in our numerical studies. It can be shown that for certain low-dimensional examples, the CI is equivalent to the two-sample t -statistic.

Similar to SigClust, we use simulation in lieu of permutation, and make use of the soft-thresholding method ([Huang et al., 2014](#)) for eigenvalue estimation. SigPal randomly labels some observations in the simulated data. This extra step is essential to mimic the true null distribution of the test statistic.

As will be shown later, although SigClust has some power when the signal within the data is relatively large, it is substantially less powerful when the signal is weak. SigPal, on the other hand, has a great power in both cases. Secondly, when the data come from a mixture of two Gaussian distributions and the mean difference is large enough, the labeled

data may not provide additional boost in the power compared to SigClust. In this case, it may make sense to simply apply SigClust and ignore the label information. As will be seen in the later sections, the strength of SigPal lies on the usefulness of the labeled data: it is visibly more powerful than SigClust when the signal is small.

4 Theoretical Property

In this section, we provide some theoretical justification for the SigPal method. We first derive the relationship between the theoretical version of the CI (TCI) and the eigenvalues of the covariance. Specifically, we assume that $\mathbf{X} \sim N(\mathbf{0}, \Sigma)$ and consider using $S^3\text{LDA}$ for class assignment. The theoretical $S^3\text{LDA}$ (Lu and Qiao, 2015) coefficient $\hat{\omega}$ is defined as

$$\begin{aligned}\hat{\omega} &= \underset{\|\omega\|=1, b=0}{\operatorname{argmin}} \quad \mathbb{E}_{(\mathbf{X}, Y)}(Y - (\omega' \mathbf{X} + b))^2 + C \mathbb{E}_{\mathbf{X}}(1 - |\omega' \mathbf{X} + b|)_+, \\ &= \underset{\|\omega\|=1}{\operatorname{argmin}} \quad \mathbb{E}_{(\mathbf{X}, Y)}(Y - \omega' \mathbf{X})^2 + C \mathbb{E}_{\mathbf{X}}(1 - |\omega' \mathbf{X}|)_+, \end{aligned}$$

where $C > 0$ is a constant. We consider the case where the effect of the unlabeled data portion dominates, that is, we let $C \rightarrow \infty$. The relationship between the theoretical cluster index (TCI) for SigPal and the eigenvalues of Σ is stated in Theorem 1.

THEOREM 1. *Suppose that $\mathbf{X} \sim N(\mathbf{0}, \Sigma)$, $P(Y = +1) = P(Y = -1) = 1/2$ and the proportion of the labeled data is θ . Assume that Σ has an eigen-decomposition $\Sigma = V' \Lambda V$, where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$. Let \mathbf{v}_1 be the first principal component direction. Then when $C \rightarrow \infty$, $\hat{\omega} = \mathbf{v}_1$, and the corresponding TCI is*

$$TCI = 1 + \theta - \frac{2}{\pi}(1 - \theta)^3 \frac{\lambda_1}{\sum_{j=1}^d \lambda_j}.$$

Theorem 1 shows that given θ , the optimal TCI only relies on the largest eigenvalue λ_1 and the sum of eigenvalues $\sum_{i=1}^d \lambda_i$. In practice, the estimations of these two quantities have a critical impact on the p -values, defined as the proportion of the CI's from the simulated

data that are less than that from the original data. In particular, let $\hat{\lambda}_i$ denote the estimate of λ_i . Assume that $\theta = 1/2$ (for a mere illustration), then the difference between the true TCI (the one for the original distribution) and the TCI resulting from a Gaussian distribution with covariance $\hat{\Lambda}$ (the estimated Λ) is proportional to,

$$E = \frac{\hat{\lambda}_1}{\sum_{i=1}^d \hat{\lambda}_i} - \frac{\lambda_1}{\sum_{i=1}^d \lambda_i}. \quad (1)$$

For hard-thresholding method, define the potential biases in the estimation of λ_1 and $\sum_{i=1}^d \lambda_i$ as δ_1 and Δ respectively. Then

$$E = \frac{\lambda_1 + \delta_1}{\sum_{i=1}^d \lambda_i + \Delta} - \frac{\lambda_1}{\sum_{i=1}^d \lambda_i} = \frac{\sum_{i=1}^d \lambda_i \delta_1 - \lambda_1 \Delta}{\sum_{i=1}^d \lambda_i (\sum_{i=1}^d \lambda_i + \Delta)}.$$

The hard-thresholding method will tend to be anti-conservative when $E < 0$, or $\lambda_1 \Delta > \delta_1 \sum_{i=1}^d \lambda_i$, that is, when the first eigenvalue is large relative to the rest (assuming that Δ is positive, which is very likely for the hard-thresholding method since the smallest eigenvalues are replaced by the background noise level.) On the other hand, the soft-thresholding method is energy preserving in the sense that the sum of the soft eigenvalues is the sum of the sample eigenvalues, and thus $\Delta = 0$. It follows that when $\hat{\lambda}_1 < \lambda_1$, that is, when the the largest eigenvalue is under-estimated, the soft-thresholding method will be anti-conservative. This happens when the first eigenvalue is only a little larger than the background noise. A detailed discussion about the results of hard-thresholding and soft-thresholding methods can be found in [Huang et al. \(2014\)](#).

We further explore the impact of the estimation of λ_i on the TCI in SigPal and SigClust. Let TCI_{SigPal} and $TCI_{SigClust}$ denote the TCI's of SigPal and SigClust respectively, and \hat{TCI}_{SigPal} and $\hat{TCI}_{SigClust}$ the TCI's from Gaussian distributions based on estimated covariance (that is, the simulated data in both SigPal and SigClust). By Theorem 3 of [Huang et al. \(2014\)](#),

$$TCI_{SigClust} = 1 - \frac{2}{\pi} \frac{\lambda_1}{\sum_{i=1}^d \lambda_i}.$$

Then when $\theta = 1/2$ (again, for an illustration), the differences are

$$TCI_{SigPal} - TCI_{SigClust} = 1/2 + \frac{7}{4\pi} \frac{\lambda_1}{\sum_{i=1}^d \lambda_i}. \quad (2)$$

$$\hat{TCI}_{SigPal} - \hat{TCI}_{SigClust} = 1/2 + \frac{7}{4\pi} \frac{\hat{\lambda}_1}{\sum_{i=1}^d \hat{\lambda}_i}. \quad (3)$$

A SigPal/SigClust test gains power if $TCI \ll \hat{TCI}$, that is, the TCI for the original data is less than the TCI from simulated data. When $\frac{\hat{\lambda}_1}{\sum_{i=1}^d \hat{\lambda}_i} > \frac{\lambda_1}{\sum_{i=1}^d \lambda_i}$, $\hat{TCI}_{SigPal} - TCI_{SigPal}$ is greater than $\hat{TCI}_{SigClust} - TCI_{SigClust}$ due to (2) and (3). Therefore, SigPal is more powerful than SigClust when $\frac{\hat{\lambda}_1}{\sum_{i=1}^d \hat{\lambda}_i} > \frac{\lambda_1}{\sum_{i=1}^d \lambda_i}$. Particularly for soft-thresholding method, $\sum_{i=1}^d \hat{\lambda}_i = \sum_{i=1}^d \lambda_i$ because it is energy preserving. When the first eigenvalue is very large relative to the others, it is easier to be over-estimated, in which case SigPal will be more powerful.

It is also of interest to look at the change of the difference $TCI_{SigPal} - TCI_{SigClust}$ with respect to θ when the eigenvalues are fixed. Assume that $\frac{\lambda_1}{\sum_{i=1}^d \lambda_i} = 1/2$, then we have the difference

$$TCI_{SigPal} - TCI_{SigClust} = \frac{1}{\pi} [\theta^3 - 3\theta^2 + (\pi + 3)\theta].$$

We plot the difference in Figure 3, which shows that the difference is almost linear in θ . It also shows that when $\theta = 0$, the difference equals 0 too. This is because when the proportion of the labeled data is 0, then all of the data are unlabeled, in which case the data set is reduced to the SigClust setting. Hence, $TCI_{SigPal} = TCI_{SigClust}$. As θ increases, labeled data play a more and more important role, and the difference of TCI_{SigPal} and $TCI_{SigClust}$ increases almost linearly with respect to θ .

In the next theorem, we study some asymptotic properties of SigPal. Since our main interest is in the HDLSS data, we choose to consider asymptotics for $d \rightarrow \infty$ with n fixed. Such kind of HDLSS asymptotic properties were previously studied by [Hall et al. \(2005\)](#), [Ahn et al. \(2007\)](#), [Liu et al. \(2008\)](#), [Jung and Marron \(2009\)](#), [Qiao et al. \(2010\)](#), [Jung et al. \(2012\)](#), [Qiao and Zhang \(2015b\)](#), among others.

Consider a mixture of two Gaussian distributions, $\eta N(\mathbf{0}, \mathbf{D}) + (1 - \eta) N(\boldsymbol{\mu}, \mathbf{D})$, where

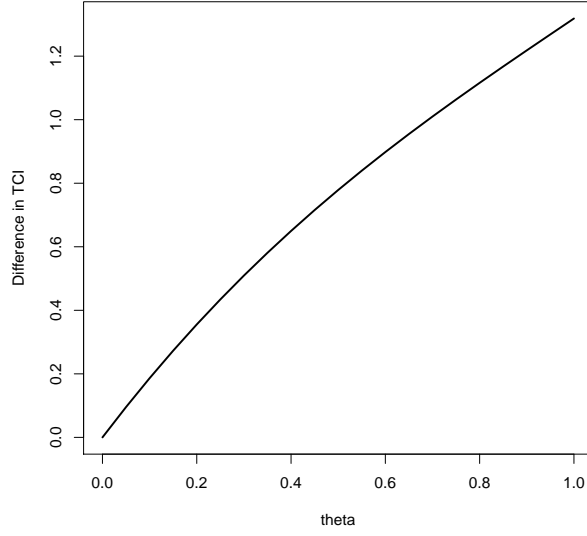


Figure 3: The difference $TCI_{SigPal} - TCI_{SigClust}$ with respect to θ , the proportion of the labeled data.

$\eta \in (0, 1)$ is the proportion for the mixture, $\boldsymbol{\mu} = (a_1, \dots, a_d)'$ a constant vector and \mathbf{D} is a diagonal matrix with diagonal elements $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$. Note that the theoretical variance for the i th variable of the mixture is $\lambda_i + \eta(1 - \eta)a_i^2$. We assume that λ_1 and a_i 's are bounded.

THEOREM 2. *Suppose that the data come from a mixture of two Gaussian distributions, $\eta N(\mathbf{0}, \mathbf{D}) + (1 - \eta)N(\boldsymbol{\mu}, \mathbf{D})$, where $\eta \in (0, 1)$, $\boldsymbol{\mu} = (a_1, \dots, a_d)'$ and \mathbf{D} is a diagonal matrix with diagonal elements $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$. Let n_1 and n_2 be the sample sizes with $\min(n_1, n_2) > 0$ and $n_1 + n_2 = n \geq 3$. Assume that $\sum_{j=1}^d \lambda_j = O(d^\beta)$ with $0 \leq \beta < 1$, $\sum_{j=1}^d a_j^2 = O(d)$, $\sum_{j=1}^d a_j^2 \lambda_j = O(d^\gamma)$ with $\gamma < 2$ and $\max_j(\lambda_j + \eta(1 - \eta)a_j^2) \leq M$ with $M > 0$ a fixed constant. If \mathbf{D} is known, then for a fixed n , the corresponding SigPal p -value converges to 0 in probability as $d \rightarrow \infty$.*

[Liu et al. \(2008\)](#) studied a similar result for SigClust in a special case when $a_1 = a_2 = \dots = a_d = a$, where a is a fixed constant. While we add some more assumptions, the theorem shows the asymptotic property for a more general setting. Theorem 2 shows that if the data come from a mixture of two Gaussian distributions, then under some assumptions, SigPal

tends to reject the null hypothesis when n is fixed and $d \rightarrow \infty$. This result justifies the usefulness of SigPal in the HDLSS data setting.

5 Numerical Study

In this section we use simulated and real examples to demonstrate the effectiveness of our proposed method.

5.1 Simulations

We use the same simulation setting as in [Liu et al. \(2008\)](#) and [Huang et al. \(2014\)](#). Three types of examples are studied, including three cases under both the null and alternative hypotheses. The sample size for all examples is $n = 40$ and dimension $d = 300$. In the first case, we consider examples of data under the null hypothesis, that is, having one cluster generated from a single Gaussian distribution. In each example, we check the type-I error by studying how often SigPal incorrectly rejects the null hypothesis H_0 . In the second and the third cases, we consider data from a collection of mixtures of two Gaussian distributions with different signal sizes and explore the power of our method in terms of how often it correctly rejects the null hypothesis. For simplicity, we consider diagonal covariance matrix \mathbf{D} because of the rotation invariance property of CI.

For each simulation, we consider two class assignment methods to be applied in SigPal, namely, S^3 LDA and COP-KMEANS. They are applied for both original data and simulated data before we calculate CI. Under the null hypothesis, theoretically the p -value should follow the Uniform $[0,1]$ distribution. Then a level α test rejects the null hypothesis with probability α when \mathbf{D} is known. This can be shown by a direct use of the standard probability integral transformation theorem. To simplify the computation, we fix the tuning parameters in S^3 LDA (the goal here is not perfect classification.) We also consider an option which uses L_1 -LDA for labeled data only on the simulated data while still using S^3 LDA on the original data. Note that this does not affect the size of the test, however, could sacrifice the power.

| v | w | L_1 -LDA | S^3 LDA | COP-KMEANS | SigClust |
|-----|----|------------|-----------|------------|----------|
| 100 | 1 | 0 | 3 | 4 | 0 |
| 50 | 2 | 1 | 4 | 3 | 0 |
| 20 | 5 | 0 | 1 | 1 | 0 |
| 10 | 10 | 0 | 2 | 2 | 0 |
| 1 | 1 | 0 | 0 | 0 | 0 |
| 3 | 1 | 0 | 0 | 0 | 0 |
| 5 | 1 | 0 | 0 | 0 | 0 |
| 10 | 1 | 0 | 0 | 2 | 0 |
| 20 | 1 | 0 | 2 | 4 | 0 |
| 50 | 1 | 1 | 3 | 4 | 0 |
| 1 | 5 | 0 | 0 | 0 | 0 |
| 10 | 5 | 0 | 1 | 2 | 0 |
| 20 | 5 | 0 | 1 | 2 | 0 |
| 50 | 5 | 0 | 2 | 1 | 0 |

Table 1: Summary table for the one cluster case over 100 replications based on different methods under different settings (different pairs of v and w). The numbers of empirical p -values which are less than 0.05 are reported.

We apply these different versions of SigPal to compare with SigClust in each case. To make the notation simple, we use L_1 -LDA to denote SigPal with S^3 LDA on the original data and L_1 -LDA on the simulated data, and use S^3 LDA and COP-KMEANS to denote SigPal with corresponding methods on both original and simulated data.

5.1.1 Case 1: One Cluster

Suppose that the data are generated from a single multivariate Gaussian distribution with covariance \mathbf{D} , where \mathbf{D} is diagonal with diagonal elements $(\underbrace{v, \dots, v}_w, 1, \dots, 1)$, that is, there are w v 's and $(d - w)$ 1's. We randomly select 20 observations to be labeled from all the 40 observations and consider 14 combinations of (v, w) as shown in Table 1. Each simulation is repeated 100 times.

In Table 1, as expected, all methods maintain the size (fewer than 5% of the p -values are less than 0.05.) L_1 -LDA is very conservative for all the 14 settings as it almost never rejects the null. A possible explanation is that under the null hypothesis, that is, when the data are generated from a single Gaussian distribution, a semi-supervised classification method

like S^3 LDA makes the class difference even smaller than applying L_1 -LDA on labeled data only (since the former attempts to incorporate the useless information.) As a consequence, the CI from the original data (after applying S^3 -LDA) is often greater than the CI's from the simulated data (after applying L_1 -LDA), and hence the p -value is often very large.

5.1.2 Case 2: Mixture of Two Gaussian Distributions with Signal in One Coordinate Direction

We now consider data generated from a mixture of two Gaussian distributions, $0.5N(-\boldsymbol{\mu}, \mathbf{D}) + 0.5N(\boldsymbol{\mu}, \mathbf{D})$, where $\boldsymbol{\mu} = (a, 0, \dots, 0)'$ and $\mathbf{D} = \text{diag}(\underbrace{v, \dots, v}_w, 0, \dots, 0)$ a diagonal matrix. The sample size is $n = 40$. From each class, we randomly choose 10 observations which we keep class labels for. Two types of the covariance matrix are conducted here, $v = 100, w = 1$ and $v = 2, w = 50$. The choices of a depend on the values of v and w . Note that when $a = 0$, the distribution reduces to a single Gaussian distribution. When $a > 0$, the population is a mixture of two Gaussian distributions and the larger the a , the greater the separation between the two distributions is. When the signal a is large enough, labeled data do not help on distinguishing the two distributions (they may even make it worse.) Thus one may ignore the label information and simply apply SigClust. In our study, we focus on the cases with small signals, in other words, when the mean difference of the two distributions is not too large. In these cases, labeled data can greatly help to gain extra power in SigPal. The empirical distributions of p -values are shown in Figure 4 and 5 for the two settings ($v = 100, w = 1$ and $v = 2, w = 50$).

Figure 4 shows the setting $v = 2$ and $w = 50$ and Figure 5 shows the spiked model setting $v = 100$ and $w = 1$. Colors are used to represent different values of a . When $a = 0$, the data are generated from a single Gaussian distribution. When $a > 0$, we study the power of the test using different methods. We consider $a = 1, \dots, 5$ for the setting $v = 2$ and $w = 50$ and $a = 5, 10, 15, 18, 20$ for the setting $v = 100$ and $w = 1$. We can see in Figure 4 that SigClust is too conservative when $a = 1, 2, 3$ and there is almost no power when $a = 1, 2$. All the three SigPal methods present more power in these settings.

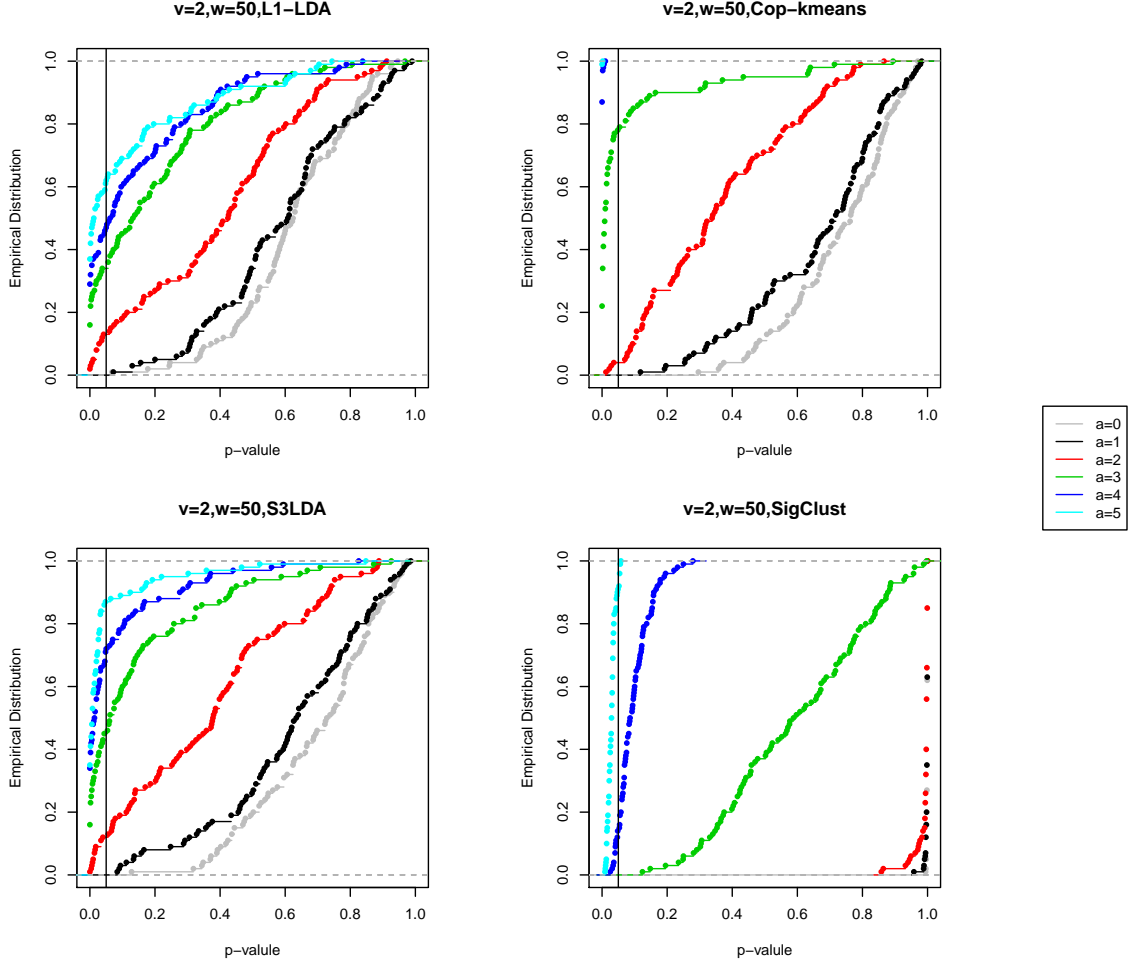


Figure 4: Empirical distributions of p -values of a mixture of two Gaussian distributions with the signal in one direction. Results are based on different methods under the setting $v = 2$ and $w = 50$, with the increase of signal a .

For the spiked model setting in Figure 5 where $v = 100$ and $w = 1$, SigClust is anti-conservative, indicated by the fact that the p -value has a higher chance of having a smaller value (the grey curve is above the 45 degree line.) On the other hand, S^3 LDA and COP-KMEANS are more powerful than SigClust. L_1 -LDA loses some power as it only uses the labeled data on the simulated data assignments. The comparison on the left two subfigures (L_1 -LDA versus S^3 LDA) also illustrates the effect of using a semi-supervised classification for label assignment compared to using a classification method. A greater power is retained as a consequence.

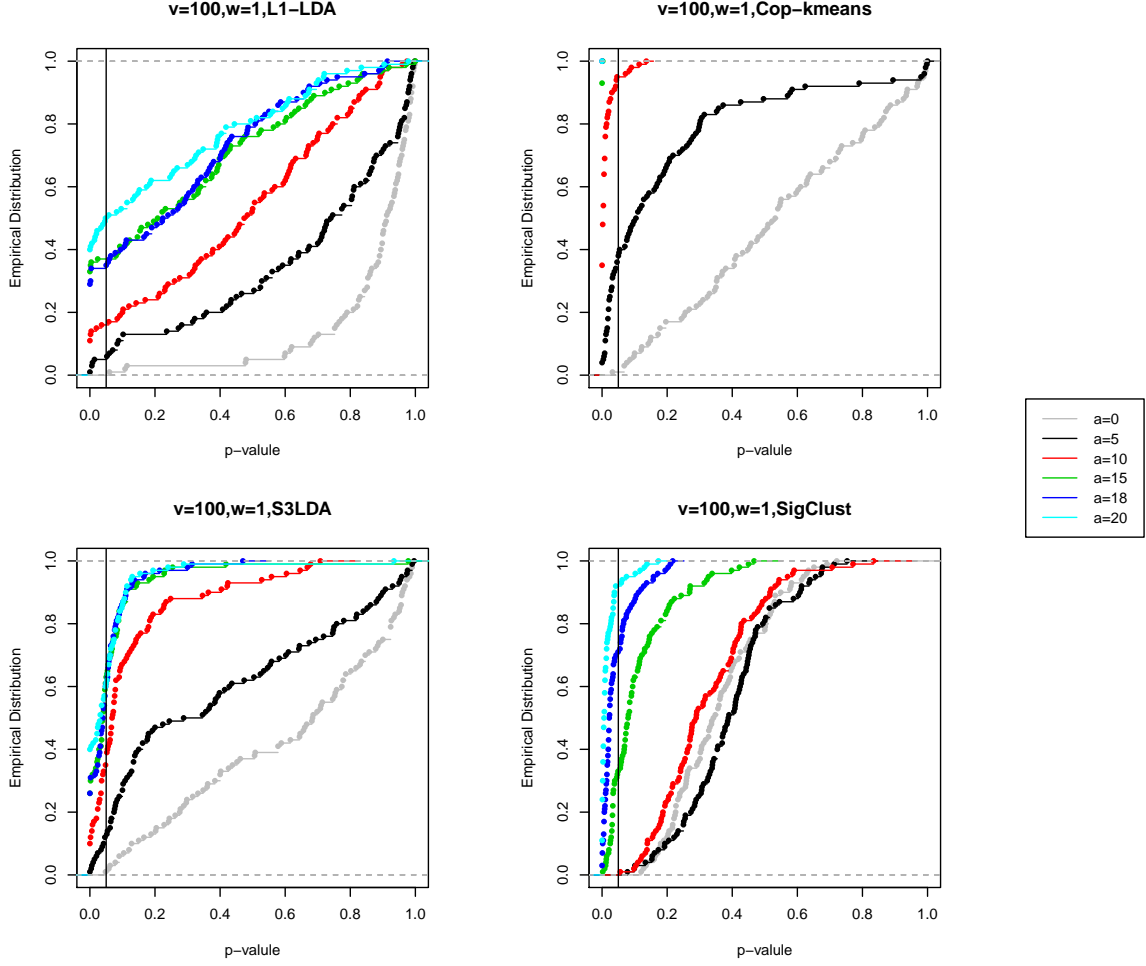


Figure 5: Empirical distributions of p -values of a mixture of two Gaussian distributions with the signal in one direction. Results are based on different methods under the setting $v = 100$ and $w = 1$, with the increase of signal a .

To make the simulation closer to the reality, we also use the Human Lung Carcinomas Microarray Dataset to obtain a more realistic covariance structure \mathbf{D} . This data set was previously analyzed in [Bhattacharjee et al. \(2001\)](#). [Liu et al. \(2008\)](#) used this data as a test bed to demonstrate their proposed SigClust approach. We extract four biological groups in the data set, 20 pulmonary carcinoid samples (Carcinoid), 13 colon cancer metastasis samples (Colon), 17 normal lung samples (normal) and 6 small cell carcinoma samples (SmallCell), with a total of 56 samples. We remove the first three principal components of the data by reconstructing the covariance matrix using the remaining terms of the eigen-expansion. The

resulting covariance \mathbf{D} is used to generate the original data, and the estimated covariance is used to generate the simulated data in the **Simulation** step of SigPal.

In terms of the signal, we consider the signal a on one direction ranging from $1, 2, \dots, 7$. The empirical distributions of p -values are displayed in Figure 6.

Figure 6 shows that all of S^3 LDA, COP-KMEANS and SigClust are strongly anti-conservative under the null hypothesis ($a = 0$), which is not the case in either Figure 4 or Figure 5. The data in Figure 6 is generated from a non-diagonal covariance matrix \mathbf{D} while Figure 4 and 5 use a diagonal covariance \mathbf{D} . Since CI is rotation invariant, it suggests that these methods become anti-conservative due to the estimation of covariance matrix.

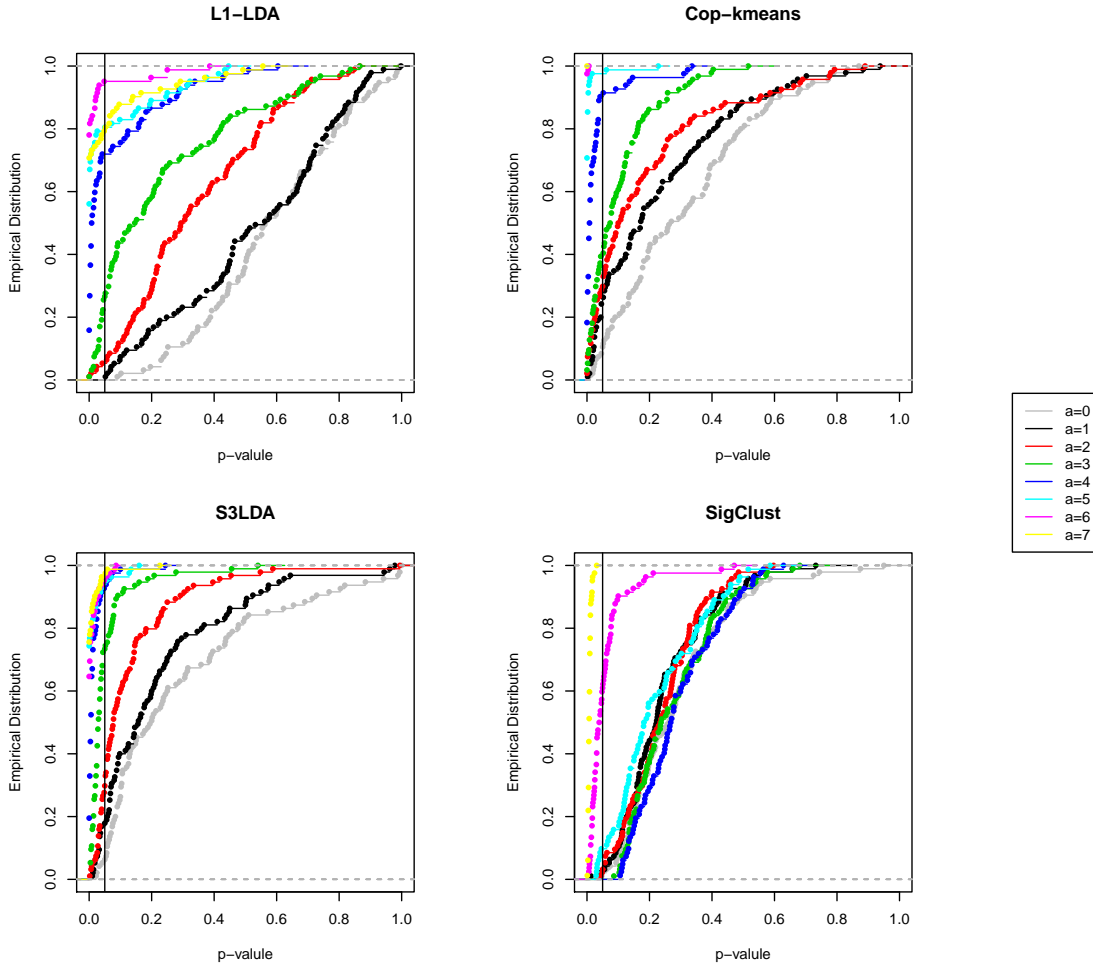


Figure 6: Empirical distributions of p -values of a mixture of two Gaussian distributions, generated by the covariance matrix from the real data, with the signal in one direction.

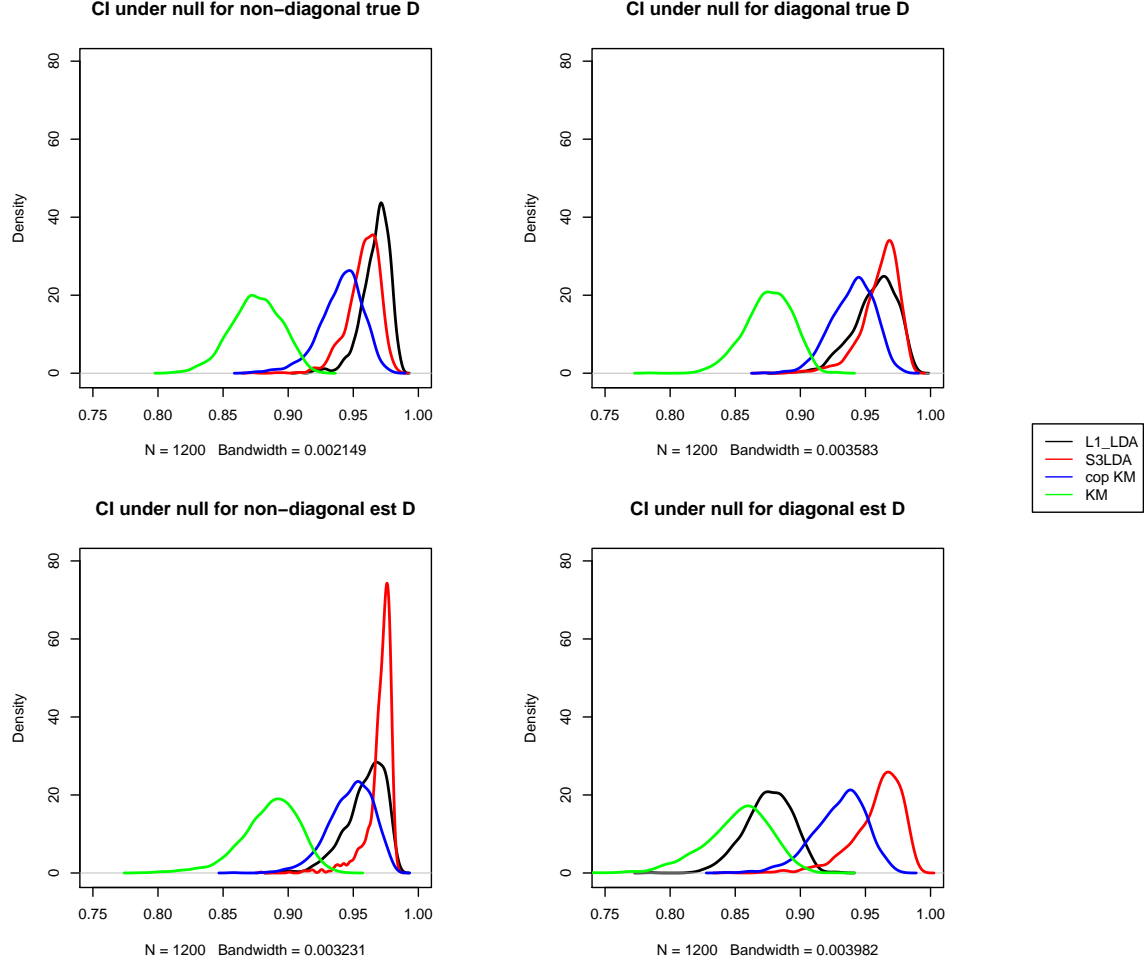


Figure 7: Understanding why our method is anti-conservative for non-diagonal covariance by comparing with the diagonal case.

To further understand the influence from the estimation of diagonal and non-diagonal covariances, we compare the distributions of CI under null hypothesis in Figure 7. The left two plots show the distributions of CI for the data generated from a non-diagonal covariance matrix (top) and for the data generated from the estimated covariance (bottom). It shows that the density curves of CI for the simulated data (bottom) for S^3 LDA, COP-KMEANS and SigClust (red, blue and green curves) are all shifted to the right compared to the case with the true distribution (top). The CI for the simulated data (bottom) is more likely to be greater than the CI for the original data (top), which makes the p -value small more often and hence the three methods become anti-conservative. This is consistent with the result

we see in Figure 6.

For the right two plots, we rotate and obtain a diagonal covariance matrix by eigen-decomposition and plot the empirical distributions of CI for the data generated from the true diagonal covariance (top) and from its estimation (bottom). The density curves of the CI for the simulated data (bottom) for S^3 LDA, COP-KMEANS and SigClust almost remain in the same position as those using the true covariance; the curve for L_1 -LDA shifts greatly to the left. Based on the comparison between the left and the right panels, we confirm our previous finding that S^3 LDA, COP-KMEANS and SigClust become anti-conservative when \mathbf{D} is non-diagonal due to the influence from covariance estimation.

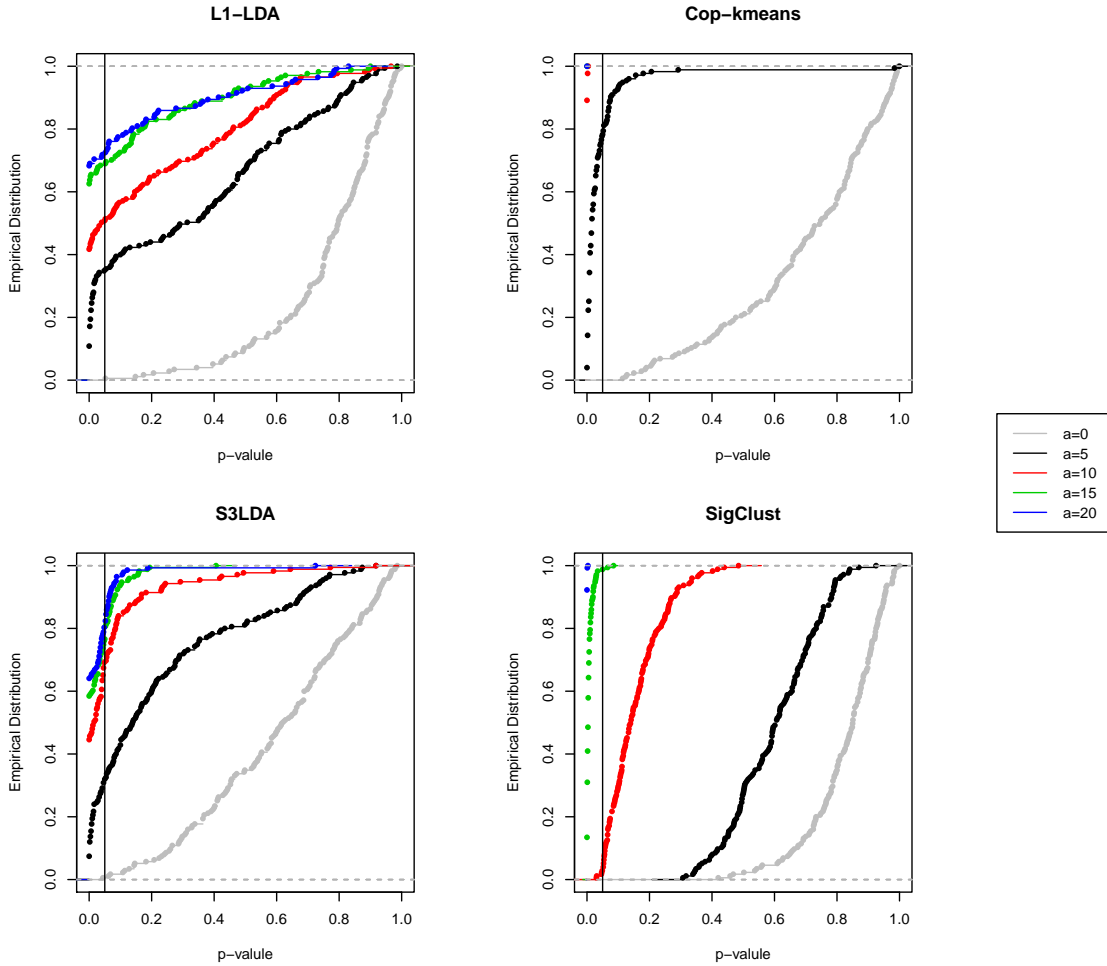


Figure 8: Empirical distributions of p -values of a mixture of two Gaussian distributions, generated by a diagonal covariance matrix from the real data, with the signal in one direction.

After conducting eigen-decomposition of the covariance matrix used in Figure 6, we obtain a diagonal \mathbf{D} and use it to generate the data. New results are presented in Figure 8. None of the methods is anti-conservative and all the three versions of SigPal are more powerful than SigClust when the signal a is relatively small ($a = 5, 10$). One is recommended to rotate the data to form a diagonal covariance matrix before applying SigPal.

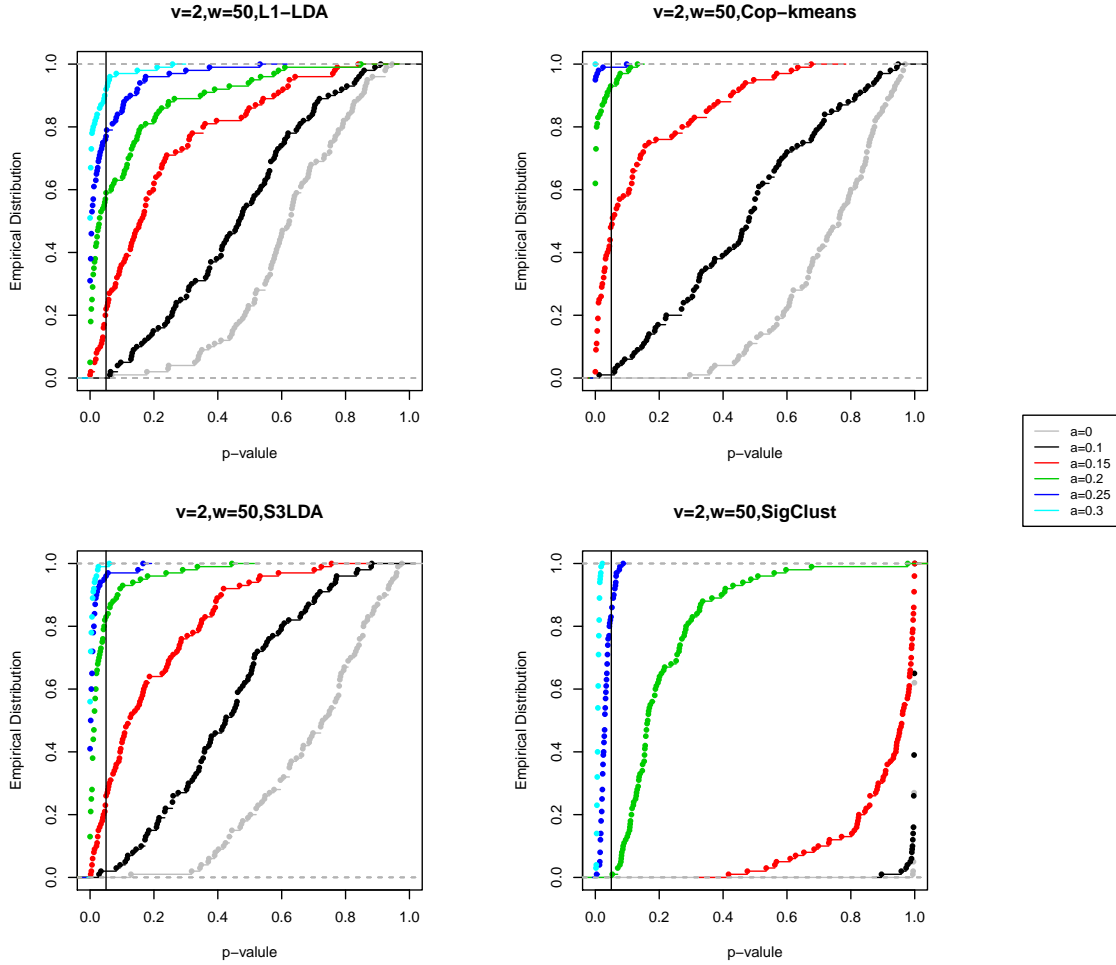


Figure 9: Empirical distributions of p -values of a mixture of two Gaussian distributions with the signal in all directions. Results are based on different methods under the setting $v = 2$ and $w = 50$, with the increase of signal a .

5.1.3 Case 3: Mixture of Two Gaussian Distributions With Signal in All Coordinate Directions

Similarly as in Figure 4, we see in Figure 9 that SigClust is too conservative when $a = 0.1$ and 0.15. All the three SigPal methods perform more powerfully than SigClust when a is less than 0.25. For the spiked model in Figure 10 where $v = 100$ and $w = 1$, SigClust is anti-conservative. SigPal is more powerful than SigClust when $a \leq 0.6$.

Now we further consider examples with signals in all coordinate directions. We generate

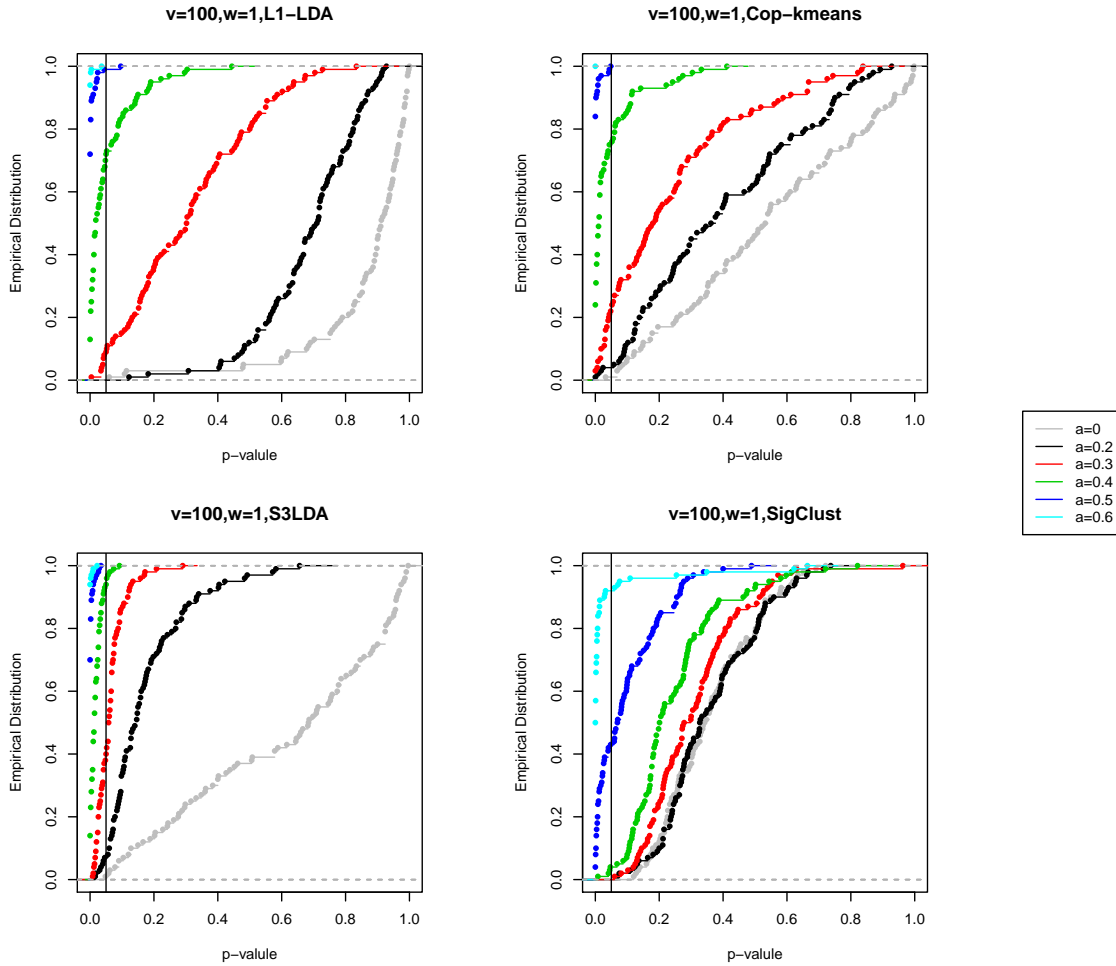


Figure 10: Empirical distributions of p -values of a mixture of two Gaussian distributions with the signal in all directions. Results are based on different methods under the setting $v = 100$ and $w = 1$, with the increase of signal a .

data from a mixture of two Gaussian distributions, $0.5N(-\boldsymbol{\mu}, \mathbf{D}) + 0.5N(\boldsymbol{\mu}, \mathbf{D})$, where $\boldsymbol{\mu} = (a, \dots, a)'$ and $\mathbf{D} = \text{diag}(\underbrace{v, \dots, v}_w, 0, \dots, 0)$ a diagonal matrix. We keep the class labels for 10 observations from each class and still consider two covariance settings, $v = 100, w = 1$ and $v = 2, w = 50$. The signal a in each direction is deliberately chosen to be very small, however, when all directions are combined together, the total signal can be very large. The empirical distributions of p -values calculated from 100 replications for the two settings are displayed in Figure 9 and 10.

In summary, SigPal maintains the size under the null distribution while SigClust is anti-conservative when the first eigenvalue is very large relative to the others. In all the cases when the signal between the two distributions is small, SigPal is relatively more powerful than SigClust due to the help from labeled data. Among the three versions of SigPal we consider in the simulation study, COP-KMEANS performs the best in most cases. When the data follows a distribution with non-diagonal covariance matrix, the test could be anti-conservative. Thus rotation of the data is recommended before SigPal is applied.

5.2 Real Data Application

In this section, we apply our method to the breast cancer data (BRCA) from The Cancer Genome Atlas Research Network, which has been studied by [Fan et al. \(2006\)](#) and [Huang et al. \(2014\)](#). The data include four subtypes: LumA, LumB, Her2 and Basal. The sample size is 348, among which there are 154 LumA, 81 LumB, 42 Her2 and 66 Basal. The number of genes used in the analysis is 4000 after filtering. For every possible pairwise combination of subclasses, we randomly select 20 observations from each class to keep the class labels and the remaining observations are treated as unlabeled. We apply SigClust, SigPal and DiProPerm to every possible pair of subclasses and report their p -values in Table 2. Here we only conduct SigPal using COP-KMEANS for class assignment in this real example. Note that these three methods are using different information. SigClust does not require label information while DiProPerm is applied to the two labeled classes. Our SigPal method is designed for partially labeled data.

| | Basal.LumA | Basal.LumB | Basal.Her2 | LumA.LumB | Her2.LumB | Her2.LumA |
|-----------|------------|------------|------------|-----------|-----------|-----------|
| SigClust | 0 | 0 | 0.009 | 0.298 | 0.537 | 0.625 |
| SigPal | 0 | 0 | 0 | 0.002 | 0.002 | 0.013 |
| DiProPerm | 0 | 0 | 0 | 0 | 0 | 0 |

Table 2: SigClust, SigPal and DiProPerm p -values for each pair of subtypes for the BRCA data. With the label information, DiProPerm can always reach significant results for all pairs. With only partial information, SigPal can reach similar conclusions.

Table 2 shows that for pairs including Basal, the p -values from all three methods are 0 which implies that Basal can be well separated from the rest. For the remaining three pairs, SigClust reports large p -values, which suggests that there is no strong evidence for them to be viewed as from two different clusters if no label information is provided. In contrast, the p -values of these three pairs for DiProPerm are all 0, indicating that each pair of two classes can be significantly separated. With the help of a small portion of the label information, SigPal gains much power to distinguish the two classes and produces very close results to DiProPerm.

To further illustrate the three methods, we use the pair of LumA and LumB as an example and display the scatter plot of the projections of the data vectors onto the first two

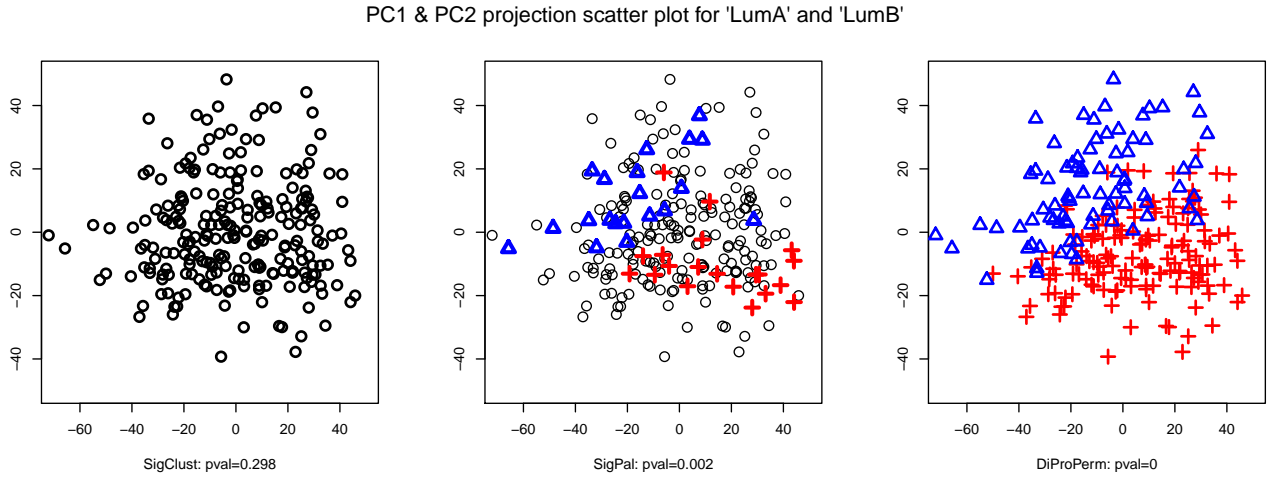


Figure 11: PCA projection scatter plot of the two classes, LumA and LumB. Colors indicate biological subtypes. LumA are displayed in red and LumB in Blue. Points in black are treated as unlabeled data. Data are analyzed by SigClust, SigPal and DiProPerm respectively in the left, middle and right panels.

principle component directions in Figure 11. Colors indicate biological subtypes. LumA are displayed in red and LumB in Blue. Points in black are treated as unlabeled data. Figure 11 shows that without given the class information (left plot), LumA and LumB seem to be one subtype so that SigClust give a non-significant result (p -value=0.298). When all the label information is available (right plot), DiProPerm suggests that these two classes are significantly different (p -value=0). With 40 labeled observation out of 235 observations in total, our SigPal method finds the difference between the two classes by extracting useful information from the small portion of the labeled data (middle plot).

6 Conclusion

In this article, we propose a significance analysis procedure, SigPal, in the HDLSS setting. This method is designed for a data set where a small amount of labeled data are available with a large amount of unlabeled data. In contrast to SigClust which does not rely on class label information, our method makes use of the labeled data to increase the difference in the classes under the null and alternative hypotheses. Through extensive simulation examples with partial label information available, we compared the performance of SigPal with SigClust in different settings. SigPal is relatively more powerful than SigClust, especially when the signal between the two classes is not large. Among the three versions of SigPal we conduct in the simulation study, COP-KMEANS performs the best in most cases.

Although CI is rotation invariant, it turns out in the simulation study that under the null hypothesis when the data comes from a distribution with non-diagonal covariance matrix, SigPal could be anti-conservative. Hence, rotation of the data is recommended before SigPal is applied.

SigPal is a general procedure with possibly many variants. The test statistic CI, used in our numerical study, may be substituted by other quantities, such as the Hotelling's T^2 statistic. There is also room for choosing different approaches to assign labels in the **Initialization** step of SigPal. An interesting and potential extension of SigPal is to case of testing multiple classes. A possible solution is to use a multi-class classification method for

the class assignment.

Appendix. Technical proofs

Proof of Theorem 1

Without loss of generality, assume that $\Sigma = \text{diag}(\lambda_1, \dots, \lambda_d)$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$. We first show that $\hat{\boldsymbol{\omega}} = \mathbf{v}_1 = (1, 0, \dots, 0)'$, which is the direction of the greatest variation of the data. Recall that $\hat{\boldsymbol{\omega}} = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmin}} \mathbb{E}_{(\mathbf{X}, Y)}(Y - \boldsymbol{\omega}'\mathbf{X})^2 + C\mathbb{E}_{\mathbf{X}}(1 - |\boldsymbol{\omega}'\mathbf{X}|)_+$. As $C \rightarrow \infty$, we only need to show that $\mathbf{v}_1 = (1, 0, \dots, 0)'$ minimizes the second term

$$\begin{aligned} & \mathbb{E}[(1 - |Z|)_+] \\ &= \mathbb{P}(|Z| \geq 1)\mathbb{E}(0|Z| \geq 1) + \mathbb{P}(|Z| < 1)\mathbb{E}(1 - |Z||Z| < 1) \\ &= 0 + \mathbb{P}(|Z| < 1)\mathbb{E}(1 - |Z||Z| < 1), \end{aligned}$$

where $Z = \boldsymbol{\omega}'\mathbf{X}$.

Let $\boldsymbol{\omega}_1 = (1, 0, \dots, 0)'$ and $\boldsymbol{\omega}_2 = (s_1, \dots, s_d)'$ with $\sum_{j=1}^d s_j^2 = 1$. Then we have $\text{Var}(\boldsymbol{\omega}_1'\mathbf{X}) = \boldsymbol{\omega}_1'\Sigma\boldsymbol{\omega}_1 = \lambda_1 \geq \text{Var}(\boldsymbol{\omega}_2'\mathbf{X})$. Let $Z_1 = \boldsymbol{\omega}_1'\mathbf{X}$ and $Z_2 = \boldsymbol{\omega}_2'\mathbf{X}$, then Z_1 and Z_2 follow Gaussian distributions with mean 0 and $\text{Var}(Z_1) \geq \text{Var}(Z_2)$. It follows that

1. $\mathbb{P}(|Z_1| < 1) \leq \mathbb{P}(|Z_2| < 1)$;
2. $\mathbb{E}(|Z_1||Z_1| < 1) \geq \mathbb{E}(|Z_2||Z_2| < 1)$.

Therefore, $\mathbb{E}[(1 - |Z_1|)_+] \leq \mathbb{E}[(1 - |Z_2|)_+]$, which implies that $\hat{\boldsymbol{\omega}} = (1, 0, \dots, 0)'$.

To show the TCI, it is enough to consider the situation with diagonal covariance matrix due to the rotation invariance of CI. We first compute the theoretical total sum of squares TSS as

$$\begin{aligned} TSS &= \mathbb{E}\|\mathbf{X}\|^2 = \int \|\mathbf{x}\|^2 \phi(\mathbf{x}) d\mathbf{x} \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \|\mathbf{x}\|^2 \phi(\mathbf{x}) dx_1 \dots dx_d \end{aligned}$$

$$\begin{aligned}
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{j=1}^d x_j^2 \left(\prod_{j=1}^d \varphi_{\lambda_j}(x_j) \right) dx_1 \cdots dx_d \\
&= \sum_{j=1}^d \int_{-\infty}^{\infty} x_j^2 \varphi_{\lambda_j}(x_j) dx_j = \sum_{j=1}^d \lambda_j,
\end{aligned}$$

where $\phi(\mathbf{x}) = \prod_{j=1}^d \varphi_{\lambda_j}(x_j) = \prod_{j=1}^d \frac{1}{\sqrt{2\pi\lambda_j}} e^{-x_j^2/2\lambda_j}$.

Recall that we assume $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and we showed $\hat{\boldsymbol{\omega}} = \mathbf{v}_1 = (1, 0, \dots, 0)'$. Here $(1, 0, \dots, 0)'$ is the norm vector of the separating hyperplane going through $\boldsymbol{\mu} = (0, \dots, 0)'$.

Let WSS be the theoretical within cluster sum of square and let WSS₁ and WSS₂ denote the theoretical sum of squares within class 1 and class 2 respectively. By symmetry the mean of class 1 $\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12}, \dots, \mu_{1d})'$ with $\mu_{12} = \mu_{13} = \cdots = \mu_{1d} = 0$. For the first dimension, note that class 1 contains the original labeled data with mean 0 with probability θ , and the original unlabeled data assigned to class 1 with mean $2 \int_0^{\infty} x_1 \varphi_{\lambda_1}(x_1) dx_1$ with probability $(1 - \theta)$, where θ is the proportion of the labeled data. Thus we have

$$\mu_{11} = (1 - \theta) \cdot 2 \int_0^{\infty} x_1 \varphi_{\lambda_1}(x_1) dx_1 + \theta \cdot 0 = (1 - \theta) \sqrt{\frac{2\lambda_1}{\pi}}.$$

So $\boldsymbol{\mu}_1 = ((1 - \theta) \sqrt{\frac{2\lambda_1}{\pi}}, 0, \dots, 0)'$. Similarly, $\boldsymbol{\mu}_2 = (-(1 - \theta) \sqrt{\frac{2\lambda_1}{\pi}}, 0, \dots, 0)'$. Then

$$\begin{aligned}
WSS_1 &= (1 - \theta) \int_0^{\infty} \cdots \int_{-\infty}^{\infty} \|\mathbf{x} - \boldsymbol{\mu}_1\|^2 \phi(\mathbf{x}) dx_1 \cdots dx_d \\
&\quad + \theta \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \|\mathbf{x} - \boldsymbol{\mu}_1\|^2 \phi(\mathbf{x}) dx_1 \cdots dx_d \\
&= (1 - \theta) \int_0^{\infty} \left(x_1 - (1 - \theta) \sqrt{\frac{2\lambda_1}{\pi}} \right)^2 \varphi_{\lambda_1}(x_1) dx_1 + (1 - \theta) \sum_{j=2}^d \int_0^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_j^2 \phi(\mathbf{x}) dx_1 \cdots dx_d \\
&\quad + \theta \int_{-\infty}^{\infty} \left(x_1 - (1 - \theta) \sqrt{\frac{2\lambda_1}{\pi}} \right)^2 \varphi_{\lambda_1}(x_1) dx_1 + \theta \sum_{j=2}^d \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_j^2 \phi(\mathbf{x}) dx_1 \cdots dx_d \\
&= \left[\frac{1}{2}(\theta + 1) + \frac{1}{\pi}(\theta^3 - 3\theta^2 + 3\theta - 1) \right] \lambda_1 + \sum_{j=2}^d \frac{1 + \theta}{2} \lambda_j.
\end{aligned}$$

Similarly, $WSS_2 = WSS_1$. Thus,

$$TCI = \frac{WSS_1 + WSS_2}{TSS} = 1 + \theta - \frac{2}{\pi}(1 - \theta)^3 \frac{\lambda_1}{\sum_{j=1}^d \lambda_j}. \quad \square$$

Proof of Theorem 2

The proof is similar to the proof of Theorem 1 in [Liu et al. \(2008\)](#). It is sufficient to show the following two points:

1. The CI ξ_1 of the data from the mixture of two Gaussian distributions, using the sources of each observation from the two Gaussian distributions as cluster assignments, converges to 0 in probability as $d \rightarrow \infty$.

2. The CI under the null hypothesis is bounded away from 0 as $d \rightarrow \infty$.

Point 1 can be shown by introducing a new data set, which is easier to work with, and a modified corresponding CI, ξ_2 . In particular, consider iid sample $\mathbf{y}_1, \dots, \mathbf{y}_n$ from $N(\mathbf{0}, \mathbf{D})$. Note that $\mathbf{x}_i \stackrel{d}{=} \mathbf{y}_i + \delta \boldsymbol{\mu}$, where $\delta = 0$ if \mathbf{x}_i comes from $N(\mathbf{0}, \mathbf{D})$ and 1 if \mathbf{x}_i comes from $N(\boldsymbol{\mu}, \mathbf{D})$. $\mathbf{x}_i \stackrel{d}{=} \mathbf{y}_i + \delta \boldsymbol{\mu}$ implies that $\xi_1 \stackrel{d}{=} \xi_2$. Let $C_{(1)}$ and $C_{(2)}$ denote the sample index sets of \mathbf{x}_i with $\delta = 0$ and $\delta = 1$ respectively. By definition, we have

$$\xi_1 = \frac{\sum_{k=1}^2 \sum_{i \in C_{(k)}} \|\mathbf{x}_i - \bar{\mathbf{x}}^{(k)}\|^2}{\sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2}.$$

ξ_2 can be written using $\mathbf{y}_1, \dots, \mathbf{y}_n$, a_1, \dots, a_d and d as

$$\begin{aligned} \xi_2 &= \frac{\sum_{k=1}^2 \sum_{i \in C_k} \|\mathbf{y}_i - \bar{\mathbf{y}}^{(k)}\|^2}{\sum_{i=1}^n \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 + \frac{n_1 n_2}{n} \sum_{j=1}^d a_j^2 + \frac{2n_1 n_2}{n} \sum_{j=1}^d a_j (\bar{y}_j^{(1)} - \bar{y}_j^{(2)})} \\ &\leq \frac{\sum_{i=1}^n \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2}{\frac{n_1 n_2}{n} \sum_{j=1}^d a_j^2 + \frac{2n_1 n_2}{n} \sum_{j=1}^d a_j (\bar{y}_j^{(1)} - \bar{y}_j^{(2)})} \\ &= \frac{d^{-1} \sum_{j=1}^d \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2}{\frac{n_1 n_2}{n} \sum_{j=1}^d a_j^2 d^{-1} + \frac{2n_1 n_2}{n} \sum_{j=1}^d a_j (\bar{y}_j^{(1)} - \bar{y}_j^{(2)}) d^{-1}}, \end{aligned} \quad (4)$$

where C_1 and C_2 denote the random grouping indices of the sample into two classes of size n_1 and n_2 , \bar{y}_j and $\bar{y}_j^{(k)}$ denote the overall sample mean and the sample mean of class k of the j th variable. Note that $\sum_{i=1}^n (y_{ij} - \bar{y}_j)^2 \sim \lambda_j \chi^2(n-1)$, $\sum_{j=1}^d \lambda_j = O(d^\beta)$ and $\beta < 1$. Thus

both the mean and variance of the numerator in (4) converge to 0 as $d \rightarrow \infty$, which implies that the numerator of (4) converges to 0 in probability.

For the denominator of (4), since $\sum_{j=1}^d a_j^2 = O(d)$, the first term converges to a constant as $d \rightarrow \infty$. For the second term, note that $\sum_{j=1}^d a_j(\bar{y}_j^{(1)} - \bar{y}_j^{(2)}) \sim N(0, \sum_{j=1}^d a_j^2 \lambda_j (\frac{1}{n_1} + \frac{1}{n_2}))$. Because $\sum_{j=1}^d a_j^2 \lambda_j = O(d^\gamma)$ with $\gamma < 2$, the second term of the denominator of (4) converges to 0 in probability as $d \rightarrow \infty$. Therefore, $\xi_2 \rightarrow 0$ in probability as $d \rightarrow \infty$, which implies $\xi_1 \rightarrow 0$ in probability as $d \rightarrow \infty$.

To show point 2, We first get the Gaussian null distribution of the mixture as $N(\mathbf{0}, \mathbf{D}^*)$, where \mathbf{D}^* is diagonal with the j th diagonal element $\lambda_j + \eta(1 - \eta)a_j^2$. Here we simply assume the null distribution with mean $\mathbf{0}$ since CI is location invariant. Let $\mathbf{z}_1, \dots, \mathbf{z}_n$ be a sample from the Gaussian null distribution. We want to show that the corresponding CI is bounded away from 0 as $d \rightarrow \infty$. To this end, we make use of the HDLSS geometry of [Hall et al. \(2005\)](#). We first check the three assumptions:

(1) the fourth moments of all entries of \mathbf{z} are uniformly bounded, by the assumption that $\max_j(\lambda_j + \eta(1 - \eta)a_j^2) \leq M$;

(2) $\lim_{d \rightarrow \infty} \text{trace}(\mathbf{D}^*)/d = \sigma^2$, where σ^2 is a constant. This is because $\lim_{d \rightarrow \infty} \text{trace}(\mathbf{D}^*)/d = \lim_{d \rightarrow \infty} \sum_{j=1}^d (\lambda_j + \eta(1 - \eta)a_j^2)/d = \eta(1 - \eta)a^2 \equiv \sigma^2$, where $a^2 = \lim_{d \rightarrow \infty} \sum_{j=1}^d a_j^2/d$;

(3) the random vector satisfies the ρ -mixing condition by the independence among the entries of \mathbf{z} .

Then it follows that $\|\mathbf{z}_i - \mathbf{z}_l\|^2 = 2d\sigma^2 + O_p(1)$, as $d \rightarrow \infty$. By the triangle inequality, we have $2(\|\mathbf{z}_i - \bar{\mathbf{z}}\|^2 + \|\mathbf{z}_l - \bar{\mathbf{z}}\|^2) \geq (\|\mathbf{z}_i - \bar{\mathbf{z}}\| + \|\mathbf{z}_l - \bar{\mathbf{z}}\|)^2 \geq \|\mathbf{z}_i - \mathbf{z}_l\|^2$ and $\|\mathbf{z}_i - \bar{\mathbf{z}}\| = \frac{1}{n} \left\| (n-1)\mathbf{z}_i - \sum_{l \neq i} \mathbf{z}_l \right\| \leq \frac{1}{n} \sum_{l \neq i} \|\mathbf{z}_i - \mathbf{z}_l\|$. Thus, as $d \rightarrow \infty$, we can bound the CI under the null hypothesis as

$$\begin{aligned} CI &= \frac{\sum_{k=1}^2 \sum_{i \in C_k} \|\mathbf{z}_i - \bar{\mathbf{z}}^k\|^2}{\sum_{i=1}^n \|\mathbf{z}_i - \bar{\mathbf{z}}\|^2} \\ &\geq \frac{\frac{1}{2}([n_1/2] + [n_2/2])2d\sigma^2 + O_p(1)}{\frac{(n-1)^2}{n}2d\sigma^2 + O_p(1)} \\ &= \frac{n([n_1/2] + [n_2/2])}{2(n-1)^2} + o_p(1), \end{aligned}$$

where $[u]$ denotes the largest integer smaller than u . This shows that under the null hypothesis CI is bounded away from 0 as $d \rightarrow \infty$. \square

References

- Ahn, J., Marron, J., Muller, K., and Chi, Y. (2007), “The high-dimension, low-sample-size geometric representation holds under mild conditions,” *Biometrika*, 94, 760.
- Bai, Z. and Saranadasa, H. (1996), “Effect of high dimension: by an example of a two sample problem,” *Statistica Sinica*, 6, 311–330.
- Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., et al. (2001), “Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses,” *Proceedings of the National Academy of Sciences*, 98, 13790–13795.
- Chandriani, S., Frengen, E., Cowling, V. H., Pendergrass, S. A., Perou, C. M., Whitfield, M. L., and Cole, M. D. (2009), “A core MYC gene expression signature is prominent in basal-like breast cancer but only partially overlaps the core serum response,” *PLoS One*, 4, e6693.
- Chapelle, O., Schölkopf, B., Zien, A., et al. (2006), *Semi-supervised learning*, MIT press Cambridge.
- Chen, S. and Qin, Y. (2010), “A two-sample test for high-dimensional data with applications to gene-set testing,” *The Annals of Statistics*, 38, 808–835.
- Collins, M. and Singer, Y. (1999), “Unsupervised models for named entity classification,” in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 189–196.
- Cortes, C. and Vapnik, V. (1995), “Support-vector networks,” *Machine learning*, 20, 273–297.

- Dempster, A. P. (1960), “A significance test for the separation of two highly multivariate small samples,” *Biometrics*, 16, 41–50.
- Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., van’t Veer, L. J., and Perou, C. M. (2006), “Concordance among gene-expression–based predictors for breast cancer,” *New England Journal of Medicine*, 355, 560–569.
- Hall, P., Marron, J. S., and Neeman, A. (2005), “Geometric representation of high dimension, low sample size data,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 427–444.
- Huang, H., Liu, Y., Yuan, M., and Marron, J. (2014), “Statistical significance of clustering using soft thresholding,” *Journal of Computational and Graphical Statistics*, 00–00.
- Jung, S. and Marron, J. (2009), “PCA consistency in high dimension, low sample size context,” *The Annals of Statistics*, 37, 4104–4130.
- Jung, S., Sen, A., and Marron, J. (2012), “Boundary behavior in high dimension, low sample size asymptotics of PCA,” *Journal of Multivariate Analysis*, 109, 190–203.
- Land, W. H., Ma, X., Barnes, E., Qiao, X., Heine, J., Masters, T., and Park, J. W. (2012), “PNN/GRNN ensemble processor design for early screening of breast cancer,” *Procedia Computer Science*, 12, 438–443.
- Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. (2008), “Statistical significance of clustering for high-dimension, low-sample size data,” *Journal of the American Statistical Association*, 103.
- Lu, Q. and Qiao, X. (2015), “Sparse Fisher’s Linear Discriminant Analysis for Partially Labeled Data,” *arXiv*, 1509.05438.
- Marron, J., Todd, M., and Ahn, J. (2007), “Distance-weighted discrimination,” *Journal of the American Statistical Association*, 102, 1267–1271.

- McLachlan, G. and Peel, D. (2004), *Finite mixture models*, John Wiley & Sons.
- McShane, L. M., Radmacher, M. D., Freidlin, B., Yu, R., Li, M.-C., and Simon, R. (2002), “Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data,” *Bioinformatics*, 18, 1462–1469.
- Qiao, X., Zhang, H., Liu, Y., Todd, M., and Marron, J. (2010), “Weighted distance weighted discrimination and its asymptotic properties,” *Journal of the American Statistical Association*, 105, 401–414.
- Qiao, X. and Zhang, L. (2015a), “Distance-weighted Support Vector Machine,” *Statistics and Its Interface*, 8, 331–345.
- (2015b), “Flexible High-dimensional Classification Machines and Their Asymptotic Properties,” *Journal of Machine Learning Research*, forthcoming.
- Sarle, W. and Kuo, A.-H. (1993), “The MODECLUS procedure,” *SAS Technical Report P-256*. SAS Institute, Cary, North Carolina.
- Schaffer, J. D., Park, J. W., Barnes, E., Lu, Q., Qiao, X., Deng, Y., Li, Y., and Land Jr, W. H. (2012), “GRNN ensemble classifier for lung cancer prognosis using only demographic and TNM features,” *Procedia Computer Science*, 12, 450–455.
- Schott, J. (2007), “Some high-dimensional tests for a one-way MANOVA,” *Journal of Multivariate Analysis*, 98, 1825–1839.
- Srivastava, M. (2007), “Multivariate theory for analyzing high dimensional data,” *Journal of the Japan Statistical Society*, 37, 53–86.
- Srivastava, M. S. and Du, M. (2008), “A test for the mean vector with fewer observations than the dimension,” *Journal of Multivariate Analysis*, 99, 386–402.
- Srivastava, M. S. and Fujikoshi, Y. (2006), “Multivariate analysis of variance with fewer observations than the dimension,” *Journal of Multivariate Analysis*, 97, 1927–1940.

- Suzuki, R. and Shimodaira, H. (2006), “Pvclust: an R package for assessing the uncertainty in hierarchical clustering,” *Bioinformatics*, 22, 1540–1542.
- Tibshirani, R. and Walther, G. (2005), “Cluster validation by prediction strength,” *Journal of Computational and Graphical Statistics*, 14, 511–528.
- Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer.
- (1998), *Statistical Learning Theory*, Wiley.
- Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., et al. (2010), “Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1,” *Cancer cell*, 17, 98–110.
- Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., et al. (2001), “Constrained k-means clustering with background knowledge,” in *ICML*, vol. 1, pp. 577–584.
- Wang, J. and Shen, X. (2007), “Large margin semi-supervised learning,” *Journal of Machine Learning Research*, 8, 1867–1891.
- Wang, J., Shen, X., and Pan, W. (2007), “On transductive support vector machines,” *Contemporary Mathematics*, 443, 7–20.
- (2009), “On efficient large margin semisupervised learning: Method and theory,” *Journal of Machine Learning Research*, 10, 719–742.
- Wei, S., Lee, C., Wichers, L., Li, G., and Marron, J. (2015), “Direction-Projection-Permutation for High Dimensional Hypothesis Tests,” *Journal of Computational and Graphical Statistics*, forthcoming.
- Wichers, L., Lee, C., Costa, D., Watkinson, P., and Marron, J. (2007), “A functional data analysis approach for evaluating temporal physiologic responses to particulate matter,” Tech. rep., Tech. Rep. 5, University of North Carolina at Chapel Hill, Department of Statistics and Operations Research.